

Improving Optical Character Recognition Performance

Mike Martin and Ann Farny

The Data Distribution Laboratory is assisting in the production of a conference CD-ROM disk for the IEEE International Geoscience and Remote Sensing Symposium (IGARSS) 94 to be held in Pasadena in August. The CD-ROM will include the following components:

- The full-text of all abstracts with full-text retrieval software.
- Scanned images of all papers for viewing with Adobe Acrobat or for printing.
- The full-text of any papers submitted electronically by the authors.
- Supplementary images or data files submitted electronically by the authors.

NCSA Mosaic, WAIS and Adobe Acrobat will be used to provide access to the contents of the CD-ROM. Acrobat provides a document reader that can be used to display and print electronic documents. The Acrobat Distiller program reads encapsulated postscript files and writes a Portable Document Format (PDF) file for viewing with the Acrobat reader program. Adobe now has a conference disk publishing package that allows conference CD-ROMs to be distributed with Mac, UNIX, PC DOS and Windows Acrobat Readers for a flat fee of \$100. All the files gathered for the CD-ROM will also be accessible via the INTERNET at the world wide web server "<http://stardust.jpl.nasa.gov/igarss>".

Over eleven-hundred hard-copy abstracts were received in January and Optical Character Recognition was performed using a Kurzweil 5200 scanner and a Xerox Imaging Systems OCR software package. For ninety percent of the abstracts, only minor editing was required on each abstract (about 3 minutes). However, about ten percent of the abstracts (120) did not OCR well. Rather than retyping these abstracts by hand, (which would take approximately 7 minutes per abstract), we decided to try an experiment utilizing two different OCR packages, Omnipage Professional for the Macintosh, and Word Scan Plus for the PC, plus a document comparison program, Docucomp II for the Macintosh. We were encouraged to try this approach due to the great success we had in a test for the Defense Nuclear Agency DARE project (Data Archival and Retrieval Enhancement). In that test we compared the results of OCR'ring four pages of scanned text and found that while there were numerous errors in each individual OCR result, the combination of the correct text from each OCR was nearly perfect.

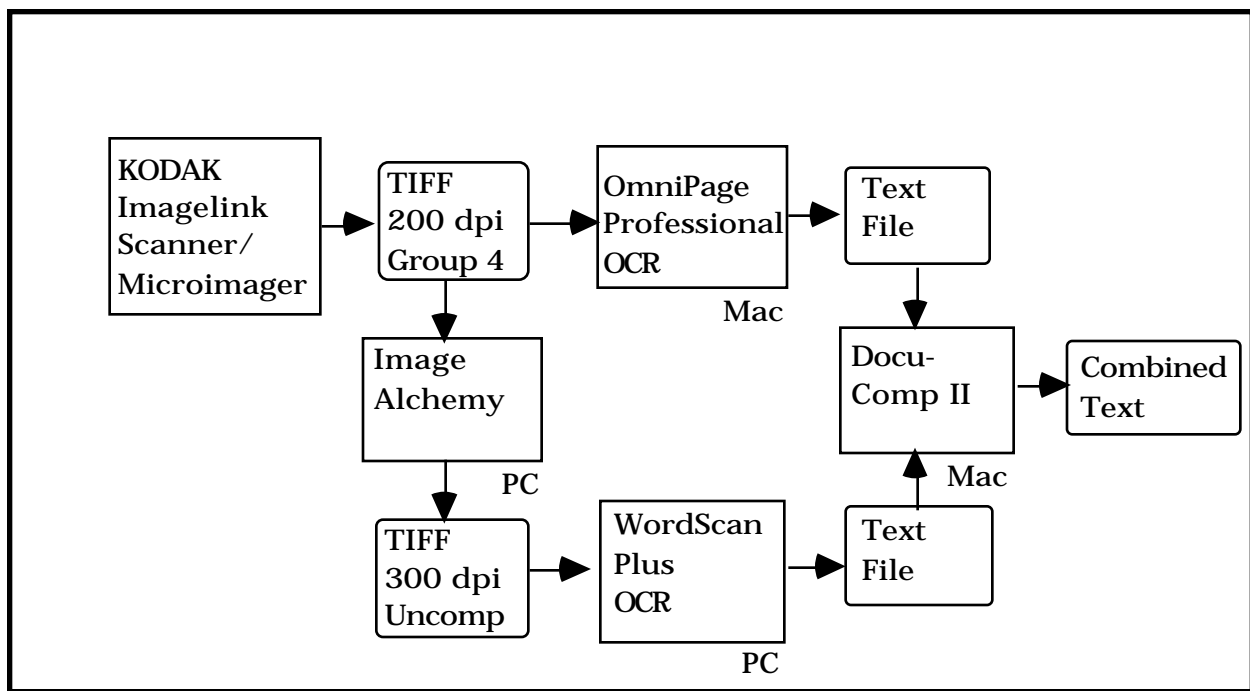


Figure 1 - Flow Diagram for Scanning, OCR and Comparison Process

Figure 1 illustrates the flow of data for the abstract processing task. The 120 papers were scanned on JPL's Kodak Scanner/Microimager 990 (which is being used by the financial records group to archive financial documents). This scanner provides 120 page-per-minute throughput and can scan both sides of a document simultaneously. The scanning took a total of about 5 minutes and the files were output as TIFF (Group 4 compression) images with 200 by 200 dots per inch resolution. The image files were then batch processed by the two OCR programs. Note the extra step on the PC to convert the image files to a format WordScan could read. It is important to note that the programs use different character recognition algorithms, which is what allows this multiple OCR technique to work. Figure 2 shows the result of processing the two scanned versions of an abstract with DocuComp II. The plain text indicates words that were identical in both scans, while the bold and underlined text show differences between the first and second document. In this paragraph the first scan has 12 errors and the second scan has 8 errors, but when the correct parts of each scan are combined only two errors remain. It is fairly easy for an editor to visually scan through such text and cut and paste to remove errors. Only rarely is it necessary to refer to the actual hard-copy abstract to determine the actual content. This allows an editor to keep eyes focused on the video display rather than going back and forth to the hard-copy. This is not only more time-consuming, but very tiring for the editor.

This double OCR technique almost guarantees that the text which is found to be identical in both OCRs is really the correct text. We did not find any examples where both OCR systems made the same recognition error.

The use of satellite images for the evaluation of structural images for the evaluation of structural geology is traditionally undertaken by geologists who manually work out lineament maps as a basis for their further interpretations. The application of the Hough transform for an automated detection of linear features have been discussed by several authors, e.g. Illingworth & Kittler, 1988; Cross & Wadge, 1988. They have demonstrated that this algorithm has the potential of detecting lines and linear features on images, and thereby allow automated mapping of lineaments for further interpretation of structural geology.

Figure 2 - DocuComp II comparison output.

It is estimated that the processing time per abstract using this method of OCR was reduced to about two minutes per abstract, saving as much as ten hours of labor over the single OCR strategy on the abstracts used in this evaluation. Remember that these abstracts were the ones that OCR'ed extremely poorly on the Kurzweil system. It is estimated that this technique could save as much as 50 percent of the effort required for average OCR results.

Lessons learned:

1. Do not rely solely on OCR'ing hard-copy. Require in your instructions to authors to submit electronically. For those who cannot comply, have the OCR process in place as a backup.
2. Make TIFF images of all material. This leaves you with complete flexibility in trying different OCR approaches.
3. Try to locate and utilize a high-speed scanner if possible. While the new scanners are expensive, they have very good page feed capability compared to older scanners.
4. Be very demanding in your instructions to authors regarding fonts, and the use of special characters or formatting. In particular, specify a

mechanism for presenting symbols and foreign letters like sigma, theta, omega.

5. Specify that names and addresses be listed sequentially. A great deal of editing is required to reformat something like the following text:

Author One	Author Two
4321 State Street	1234 City Avenue
Yourcity, St 94321	Mycity, St 80321

Which becomes the following when after OCR processing:

Author One Author Two
4321 State Street 1234 City Avenue
Yourcity, St 94321 Mycity, St 80321

5. Use a spelling checker as you edit. It will fix a lot of simple errors.

Problems.

Not everything went smoothly during this process. In particular Word Scan Plus could not read 200 dpi images, nor could it read TIFF compressed images. Thus all the image files had to be converted to 300 dpi in uncompressed format before the program could process the images. This was a trivial process using the Image Alchemy program (Handmade Software, 800-358-3588) which can read and convert nearly any type of image file in existence (it even reads the images in the PDS CD-ROM collection). A second minor problem was that the output of both OCR packages produced older versions of word processor formats which were not compatible with DocuComp II, requiring an extra step of converting to Microsoft Word 5 format before doing the document comparison.